Investment - Case Study

Aim : The objective is to identify the best sectors, countries, and a suitable investment type for making investments for Spark Funds considering the investments in range of 5 to 15 million USD in English speaking countries.

The Data given contains information about the various countries with their respective funding round types and amount raised in USD. The main objective as a business analyst is to find the appropriate sector and investment type to invest money so as to get maximum profit in a country for Spark Funds. We have given three data files in which the record is kept. We will have to analyse those files for the proper insights of the data.

The first file named Companies contains information about the companies i.e. link of the company, name , category, status, country code , funding round type, amount raised in USD etc.
Second file named round2 also contains the same information that may be means some has collected data for the second time to insure its validity.On the other hand third file named mapping contains information about category list and its main sectors. In order to obtain the final results we will have to map the three files using merge and thus get the required insights.

Code Analysis:

First of all let us import all the necessary libraries and packages that could be required for the operation.

```
import pandas as pd  #for using pandas function
import matplotlib.pyplot as plt #for plotting the data
import seaborn as sns #for plotting the data
```

After importing libraries let us import the data from the url given :

```
company=pd.read_csv(company_url,sep='\t',encoding='ISO-8859-1')
rounds2=pd.read_csv(rounds2_url,sep=',',encoding='ISO-8859-1')
```

While importing the data we were getting an error of :

```
UnicodeDecodeError: 'utf-8' codec can't decode byte 0xa0 in position 25: invalid start byte
```

Therefore we here uses one of the mostly used encoding techniques *encoding='ISO-8859-1.*

After importing the datasets it's time to do some Exploratory Data Analysis on the data sets.

```
#to observe columns in data
company.columns
rounds2.columns
#to observe the first few datasets in each of them
company.head()
rounds2.head()
```

```
#to observe the number of rows and columns in the file
company.shape
rounds2.shape
#to observe the null values
company.isna().sum()
rounds2.isna().sum()
```

After the above code there is high number of NaN cells in the datasets . Therefore for complete data analysis they should be treated properly. Here comes the part of Data Cleaning and Variable reduction.

```
#Get list of companies which are there in rounds2 but not in
company dataframe
rounds2[~rounds2['company_permalink'].isin(company['permalink'])]
```

```
#to join the companies we have to perform merging in both the
dataframes on 'company_permalink' :
#for performing the join the column must be same
rounds2.rename(columns={'company_permalink':'permalink'},
inplace=True)
rounds2.columns
master_dataframe = company.merge(rounds2, on = 'permalink', how =
'inner')
```

Now analysing the master_dataframe :

```
master_dataframe.shape
master_dataframe.info()
master_dataframe.head()
master_dataframe.columns
```

For the analysis, the non-required columns must be dropped off from the master_dataframe in order to carry out analysis:

```
#Now removing useless columns
master_dataframe.drop(['homepage_url','founded_at','funding_round_code','state_code','region','city'],inplace = True,axis = 1)
```

Now checking and removing the null values from the rows and columns.

```
#Now Treating the Null values in DataFrame
master_dataframe.permalink.isnull().sum()
master_dataframe.name.isnull().sum()
```

```
#since the value cannot be imputed in this therefore dropping this row
index=master_dataframe[master_dataframe.name.isnull()].index
master_dataframe.drop(index,inplace=True)
```

```
#removing rows with null values greater than 3
master_frame=master_dataframe[master_dataframe.isnull().sum(axis=1)<=3]
```

#removing the rows having null values in raised_amount_usd

```
master_frame=master_frame[~master_frame.raised_amount_usd.isnull()]
master_frame.raised_amount_usd.isnull().sum()
master_frame.shape
master_frame.columns
```

#treating the rows having null values in category_list

```
master_frame.category_list.isnull().sum()
```

Now naming new Dataframe as master_frame

```
master_frame= master_frame[~master_frame.category_list.isnull()]

master_frame.country_code.isnull().sum()
master_frame= master_frame[~master_frame.country_code.isnull()]

master_frame.isnull().sum()
master_frame.info()
```

Now Since there is a loss of data let us calculate how much data is retained after the reduction.

```
#retained data
retained_data= round(100*len(master_frame.index)/114942,2)
retained_data
```

The retained data comes out to be 77.02% of the original data.

Now finding the best funding type with most number numbers of investments:

```
master_frame.columns#to observe the columns name
```

```
master_frame.groupby('funding_round_type')['raised_amount_usd'].count()
```

#**Venture** is observed to be the best Funding Type

```
venture_data=master_frame[master_frame['funding_round_type']=='venture']
venture_data
```

Now finding the top9 countries investing in this funding_round_type:

```
#top 9 countries within the venture_data
countries = venture_data.groupby('country_code').sum()
top9 = countries.sort_values(by='raised_amount_usd', ascending = False).head(9)
top9
```

Identifying top 3 english speaking countries with maximum investment.

```
#Identifying top3 countries speaking English with top investment
venture_data.head()
```

```
top3=venture_data.loc[(venture_data['country_code']=='IND')              |
(venture_data['country_code']=='USA')                                    |
(venture_data['country_code']=='GBR')]
top3.head(10)
```

#Now splitting the category list based on delimiter and added into new column in top3 dataframe

```
category_data = top3['category_list'].apply(lambda x: x.split('|')[0])
category_data
top3['primarysector'] = category_data
top3.head(10)
top3.primarysector
top3.columns
```

Now getting the mapping data to map it to the top3 dataframe.

```
mapping_url='https://raw.githubusercontent.com/akjadon/Finalprojects
_DS/master/Bank%20Loan%20Default%20-%20Casestudy/mapping.csv
'
mapping_data = pd.read_csv(mapping_url, sep = ',')
mapping_data.info()
mapping_data.head()
mapping_data.columns
```

#Now converting wide dataframe into long dataframe using melt function:

```
variables = mapping_data.columns[1:]
values = mapping_data.columns[:1]
values,variables
```

```
mapping_long=pd.melt(mapping_data,id_vars=list(values),value_vars=
list(variables),var_name='main_sector', value_name='Count')
mapping_long.columns
mapping_long.rename(columns={'category_list':'primarysector'},inplac
e = True)

mapping_long = mapping_long[mapping_long.Count == 1]
mapping_long
```

Now since the two dataframe are ready we should merge the two frames on any common column, here we are going to merge it on the column named primary sector and creating a new dataframe namely dataframe

```
dataframe = pd.merge(top3, mapping_long,how = 'inner',
on='primarysector')
dataframe
```

Since we have the dataframe with all the required columns and data we should now collect the data country wise:

#Now getting the data of US based funding types with raised amount between 5 to 15 million USD

```
us_data = dataframe.loc[(dataframe.raised_amount_usd >= 5000000)
& (dataframe.raised_amount_usd <= 15000000) &
(dataframe.country_code=='USA')]
```

```
us_data.head()
```

#Now getting the data of The Great Britain based funding types with raised amount between 5 to 15 million USD

```
gbr_data = dataframe.loc[(dataframe.raised_amount_usd >= 5000000) & (dataframe.raised_amount_usd <= 15000000) & (dataframe.country_code=='GBR')]
```

```
gbr_data.head()
```

#Now getting the data of India based funding types with raised amount between 5 to 15 million USD

```
ind_data = dataframe.loc[(dataframe.raised_amount_usd >= 5000000) & (dataframe.raised_amount_usd <= 15000000) & (dataframe.country_code=='IND')]
```

```
ind_data.head()
ind_data.columns
```

Now performing the analysis on the top3 country wise data:

#performing analysis on USA data

```
us_data.head()
```

```
#total investment that took place
us_data.groupby('main_sector')['raised_amount_usd'].describe().sum().sum()
```

```
#total number of investment
us_data.groupby('main_sector')['Count'].describe().sum().sum()
```

```
#Top sectors for investments
us_data.groupby('main_sector')['raised_amount_usd'].describe()
```

```
#it is clear from above that others sector is 1st top investing sector,Cleantech / Semiconductors is 2nd
#Company receiving highest investment for top sector ('Others')
us_data[us_data['main_sector']=='Others'].groupby('name')['raised_amount_usd'].sum().sort_values(ascending=False).head()
```

```
#Company receiving highest investment for 2nd Top Sector Cleantech / Semiconductors
us_data[us_data['main_sector']=='Cleantech / Semiconductors'].groupby('name')['raised_amount_usd'].sum().sort_values(ascending=False).head()
```

```
#Performing analysis on The great britain data

gbr_data.head()
#total investment that took place
```

```
gbr_data.groupby('main_sector')['raised_amount_usd'].describe().sum().sum()
#total number of investment
gbr_data.groupby('main_sector')['Count'].describe().sum().sum()
#Top sectors for investments
gbr_data.groupby('main_sector')['raised_amount_usd'].describe()
#It is clear from above that others sector is 1st top investing sector,Cleantech / Semiconductors is 2nd
#Company receiving highest investment for top sector ('Others')
gbr_data[gbr_data['main_sector']=='Others'].groupby('name')['raised_amount_usd'].sum().sort_values(ascending=False).head()

#Company receiving highest investment for 2nd Top Sector Cleantech / Semiconductors
gbr_data[gbr_data['main_sector']=='Cleantech / Semiconductors'].groupby('name')['raised_amount_usd'].sum().sort_values(ascending=False).head()


#Performing analysis on India data

ind_data.head()

#total investment that took place
ind_data.groupby('main_sector')['raised_amount_usd'].describe().sum().sum()
```

#total number of investment

```
ind_data.groupby('main_sector')['Count'].describe().sum().sum()
```

#Top sectors for investments

```
ind_data.groupby('main_sector')['raised_amount_usd'].describe()
```

#it is clear from above that others sector is 1st top investing sector News, Search and Messaging is 2nd

#Company receiving highest investment for top sector ('Others')

```
ind_data[ind_data['main_sector']=='Others'].groupby('name')['raised_amount_usd'].sum().sort_values(ascending=False).head()
```

#Company receiving highest investment for 2nd Top Sector Cleantech / Semiconductors

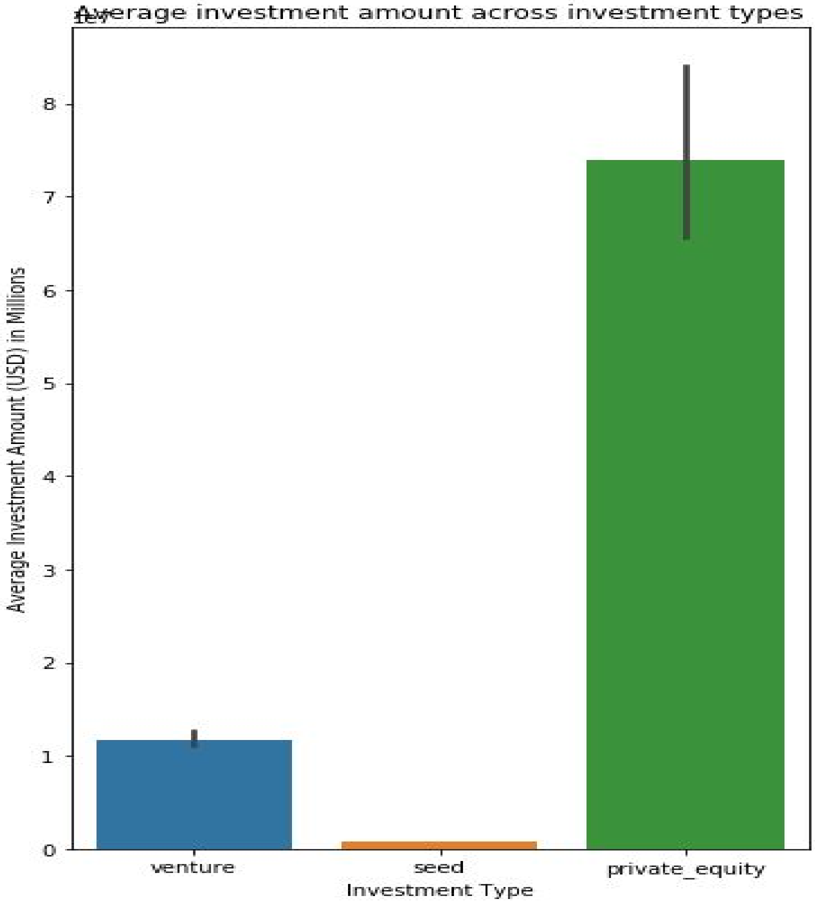
```
ind_data[ind_data['main_sector']=='News, Search and Messaging'].groupby('name')['raised_amount_usd'].sum().sort_values(ascending=False).head()
```

Now Since all the top data is collected from the top countries its time to plot the data to get the clear insights of it which is gonna help the company in getting the idea for investing.

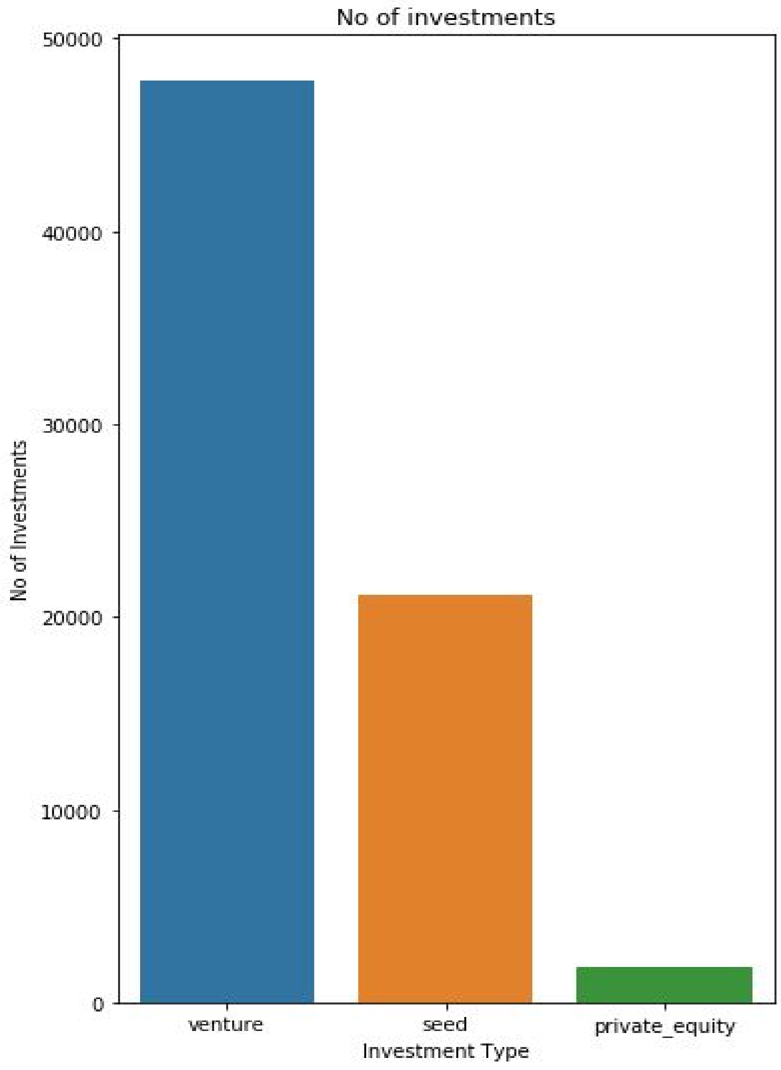
#Box and count plot to show average investments and number of investments

```
filtered_df=master_frame.loc[(master_frame['funding_round_type']=='
venture') | (master_frame['funding_round_type']=='seed') |
(master_frame['funding_round_type']=='private_equity')]

# subplot 1: Mean
plt.subplots(figsize=(20,10))
plt.subplot(1, 3, 1)
axis_bar= sns.barplot(x='funding_round_type', y='raised_amount_usd',
data=filtered_df)
axis_bar.set(xlabel='Investment Type', ylabel='Average Investment
Amount (USD) in Millions')
plt.title('Average investment amount across investment types')
```

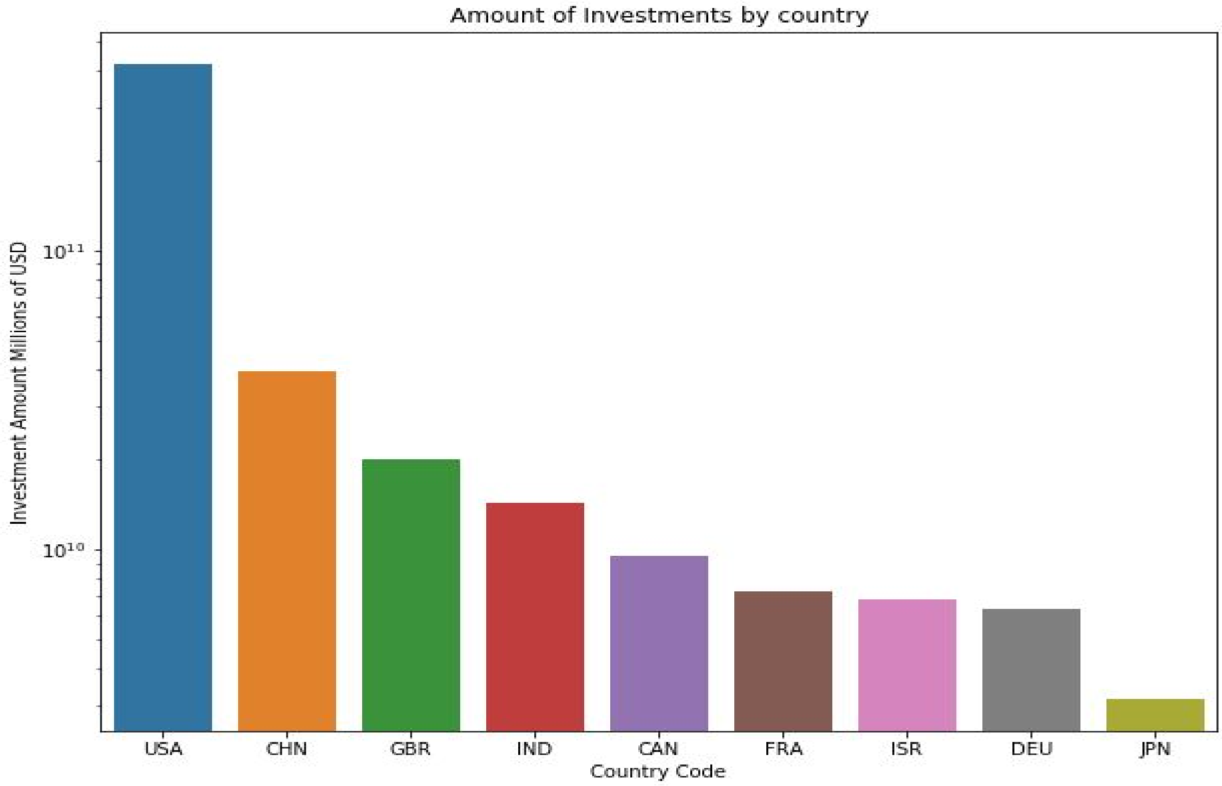


```
# subplot 2: No of investments
plt.subplots(figsize=(20,10))
plt.subplot(1, 3, 3)
axis_count = sns.countplot(x='funding_round_type' , data=filtered_df)
axis_count.set(xlabel='Investment Type', ylabel='No of Investments')
plt.title('No of investments')
plt.show()
```



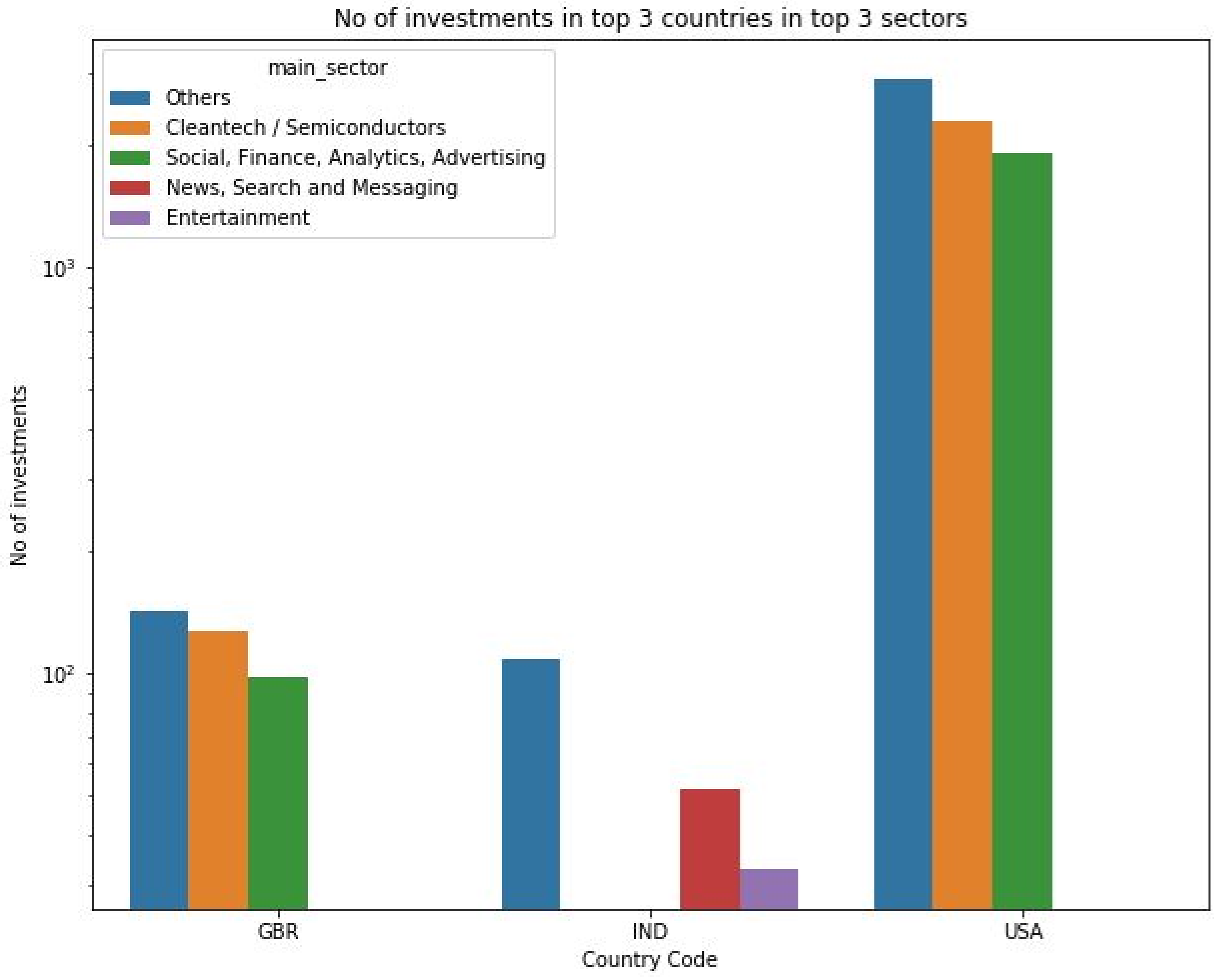
```
#Get the data frame consisting of top 9 countries investment
top9_df=pd.DataFrame({'country_code':top9.index,'raised_amount_usd
':top9.raised_amount_usd})
top9_df

#Bar plot for total no of investments by country code
plt.subplots(figsize=(10,8))
axis = sns.barplot(x='country_code', y='raised_amount_usd',
data=top9_df)
plt.yscale('log')
axis.set(xlabel='Country Code', ylabel='Investment Amount Millions of
USD')
plt.title('Amount of Investments by country')
plt.show()
```



```
#Get all the sectors with investment range between 5 to 15 million
sector_df = dataframe.loc[(dataframe.raised_amount_usd >= 5000000) & (dataframe.raised_amount_usd <= 15000000)]

#Get the investments in the main sectors by the country code
top3_df = sector_df['main_sector'].groupby(sector_df['country_code']).value_counts()
top3_df
#Get the top 3 sectors in top 3 countries in a dataframe
top3_filtered = pd.DataFrame({'count' : top3_df.groupby('country_code').head(3)}).reset_index()
top3_filtered
#Plot to show the no of investments in top 3 countries in top 3 sectors
plt.subplots(figsize=(10,8))
axis = sns.barplot(x='country_code', y='count', hue='main_sector', data=top3_filtered)
plt.yscale('log')
axis.set(xlabel='Country Code', ylabel='No of investments')
plt.title('No of investments in top 3 countries in top 3 sectors')
plt.show()
```



Hence the analysis is completed and we have the required data of the best sectors, countries, and a suitable investment type for making investments for Spark Funds considering the investments in range of 5 to 15 million USD in English speaking countries.